Available online at www.sciencedirect.com
ScienceDirect
Talanta 75 (2008) 1203–1207
Talanta
www.elsevier.com/locate/talanta

# The use of FT-NIR for API content assay in organic solvent media: A single calibration for multiple downstream processing streams

Licínia O. Rodrigues [a,*], Joaquim P. Cardoso [b], José C. Menezes [a]

[a] *IBB, Centre for Biological and Chemical Engineering, IST, Technical University of Lisbon, Avenue Rovisco Pais, P-1049-001 Lisbon, Portugal*
[b] *CIPAN S.A. Vala do Carregado, P-2601-906 Alenquer, Portugal*

Received 22 September 2007; received in revised form 7 January 2008; accepted 9 January 2008
Available online 17 January 2008

**Abstract**

The use of near infrared spectroscopy (NIRS) in downstream solvent based processing steps of an active pharmaceutical ingredient (API) is reported. A single quantitative method was developed for API content assessment in the organic phase of a liquid–liquid extraction process and in multiple process streams of subsequent concentration and depuration steps. A new methodology based in spectra combinations and variable selection by genetic algorithm was used with an effective improvement in calibration model prediction ability. Root mean standard error of prediction (RMSEP) of 0.05 in the range of 0.20–3.00% (w/w) was achieved. With this method, it is possible to balance the calibration data set with spectra of desired concentrations, whenever acquisition of new spectra is no longer possible or improvements in model's accuracy for a specific selected range are necessary. The inclusion of artificial spectra prior to genetic algorithms use improved RMSEP by 10%. This method gave a relative RMSEP improvement of 46% compared with a standard PLS of full spectral length.


*Keywords:* Near infrared; Downstream processing; Organic media; Multivariate calibration

## 1. Introduction

Lab based methods used by the pharmaceutical industry are often time consuming and add to the manufacturing cycle time. NIR spectroscopy offers one possible option for process based analytical measurements and has a significant advantage over some other technologies as it allows fast analytical measurements with simple sample preparation. However, the main feature of NIR of being sensitive to both chemical and physical effects can sometimes be a challenge when one is dealing with large sample matrix variations and very low analyte concentrations. These two conditions, which are a commonplace inside the pharmaceutical manufacturing, hinder not only the development of robust calibrations but also the application of the same models in more than one point of the process.

To overcome this problem in the present study, a new methodology was applied to build one robust calibration model capable of accurately predicting active pharmaceutical ingredient (API) content in solvent samples independently of their origin. The model robustness is based on spectral selectivity for the analyte of interest and careful wavenumbers selection.

To the best of our knowledge, the use of the same calibration for measuring streams from different points of consecutive downstream processing steps has not yet been reported.

### 1.1. Process overview

An API purification process includes a series of steps: at the end of the API fermentation, the culture media is first clarified by filtration or centrifugation and the biomass discarded. After the clarification and concentration steps, the active compound is extracted from the aqueous media to an organic phase, by liquid–liquid extraction. The extraction phase is then concentrated by evaporation. At this point, it has a cloudy orange appearance given by the degradation products. Next, to remove these non-desired products, the solution is submitted to purifying steps, where it loses most impurities but the water content increases.

* Corresponding author. Tel.: +351 21 841 9838; fax: +351 21 841 9197.
*E-mail address:* bsel@ist.utl.pt (L.O. Rodrigues).

0039-9140/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2008.01.013

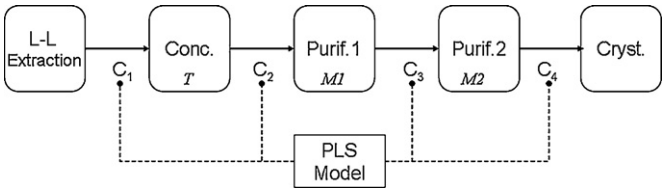


Fig. 1. Process description. The use of one single calibration model along the process is proposed.

A schematic representation of the process is presented in Fig. 1. The use of a single model along the process is proposed for assessment of concentrations C1, C2, C3 and C4.

## 2. Methods

### 2.1. Feature selection for partial least squares modelling

Partial least squares [1] is used in spectroscopy to extract relevant information from complex spectra containing overlapping absorption peaks, interferences from light scatter and noise. PLS was first considered as being almost insensitive to noise, and therefore, no feature selection was required [2]. In the last years, it has been widely recognised that variable selection can improve the prediction ability [3].

In near infrared spectroscopy (NIRS) the number of measured variables is very large and in most cases not all of them contain useful information. Theoretically, the regression coefficient for such variables will be close to zero and the inclusion of these variables should not affect negatively the calibration. However, there are significant improvements in the prediction accuracy of a calibration when a proper selection of variables is made. Moreover, the variable selection usually leads to a reduction of latent variables in the model and improve the stability of the calibration by the reduction of multicolinearity between variables [4].

Genetic algorithms (GA) have been successfully used as a method to select the most informative variables [5–7]. GAs are based upon the principle of natural evolution and selection: reproduction, mutation and selection based on fitness. The main risk of the application of GA is overfitting [5]. Meaningless variables can be selected due to the presence of non-causal correlations [8]. As such, it is necessary that the model defined neither over-fits nor under-fits the data. The selection criteria can be based on root mean squared error of cross validation (RMSECV) or on root mean squared error for prediction of an external data set (RMSEP) defined as follows:

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(Y_{i(\mathrm{cv})} - Y_{i(\mathrm{ref})})^2}{n}} \quad (1)$$

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(Y_{i(\mathrm{pred})} - Y_{i(\mathrm{ref})})^2}{n}} \quad (2)$$

where $n$ stands for the number of prediction samples, $Y_{i(\mathrm{cv})}$ and $Y_{i(\mathrm{pred})}$ for predicted values for cross validation and external validation, respectively, and $Y_{i(\mathrm{ref})}$ for reference value of sample $i$.

### 2.2. Combined spectra method (CSM)

The proposed methodology combines the spectra of two samples of different concentrations and assumes that the concentration of the averaged spectrum is also the average of the original concentration values (Fig. 2). This assumption pretends to reveal the most linearly correlated wavenumbers, assuming Beer's Law for transmission, which states that observed absorbance at a given wavelength for a mixture of different components is additive and the individual contributions are linear.

It is thus expected that proportional relations between the new spectra and the new concentrations will only be achieved for the variables truly linearly correlated with the analyte, enhancing the "good" and the "bad" spectral ranges.

Combined spectra method was performed with the following steps:

(1) Find the optimum number of latent variables and the best pre-processing method for a full PLS model with the raw calibration set (Data Set 1).
(2) Perform the spectra combination in the unprocessed spectra, for the desired concentration range. Use the Matlab function code below to create the new spectra, in which ***X*** represents the spectral matrix and ***Y*** the corresponding concentration vector, taken from Data Set 1:

```
function [xc,yc] = combspec(X,
Y,ncomb)
mix = zeros(1,size(x,2)+1);
for i = [1:10]
sx = shuffle([x y]);
mx = meanspc(sx,ncomb); mix = [mix;mx];
end
mix(1,:) = [];
yc = mix(:,end);xc = mix(:,1:end-1);
```

(3) Add the resulting spectra to the original data set and pre-process it again to build Data Set 2.
(4) Run the GA in Data Set 2.
(5) Cut Data Set 2 according to GA best fit results to obtain Data Set 3.
(6) Build the PLS model with Data Set 3 and predict an independent data set for validation.
(7) Compare resulting RMSEP with the previously RMSEP calculated in step one. If no significant improvement is reached, the algorithm should return to step 2 and new set of spectra combinations should be added to the previous one.

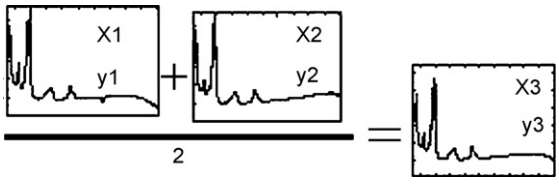


Fig. 2. Combined spectra method. It combines the spectra of two (or more) samples of different concentrations and produces an averaged spectrum and its virtual concentration.